I.S. Mayorov[1,2]
O.A. Tarasova[1]

INCREASING OF QSAR MODEL ACCURACY THROUGH DATA HOMOGENISATION ON A CASE STUDY FOR HIV-1 REVERSE TRANSCRIPTASE

[1] Institute of Biomedical Chemistry, Pogodinskaya Str., 10, Bldg. 8, 119121, Moscow, Russia;
[2] Pirogov Russian National Research Medical University, Ostrovitianov str. 1, 117997, Moscow, Russia;
mayorovmsk@gmail.com

Reverse transcriptase (RT) is one of the primary targets for antiretroviral therapy. At first, it is related to the lower speed of resistance development for reverse transcriptase inhibitors (RTI) comparing to other classes of antiretroviral drugs [1]. Secondly, RTIs act in early stages of interaction between virus particle and host cell helping to prevent virus replication [2].

Computational methods of drug development became very important for the search of new drugs [3] including antiretrovirals [4]. Building models with those methods commonly requires usage of samples taken from commercially or publicly available databases (for example ChEMBL). However, data on biological activity varies even within one database especially if different testing protocols were used. It can cause decreasing of computational model accuracy and make difficult to estimate the model's predictivity.

We have built quantitative structure-activity relationships (QSAR) models for HIV-1 RTIs using data on the biological activity of compounds taken from commercially available Thomson Reuters Integrity and freely accessible ChEMBL databases. QSAR model prediction accuracy was closely related to the way how the data sets were formed. Samples combined based on the only target, and quantitative characteristic of biological activity information showed poor accuracy. The data sets prepared taking into account the type of assay (similar biological material and method of experimental testing used to determine the biological activity of compounds) lead to more accurate models.

The purpose of our study is to develop and validate method that will allow creating homogeneous sample suitable for building accurate and predictive QSAR models. Over 150 relevant scientific publications were chosen from PubMed database. Then, their contents were converted from PDF to text format with using PDFLib TET tool. Obtained text was processed through Lingpipe 4.1.0 Java libraries and Python 2.6 scripts to select text parts containing biological assay description, prepare data sets containing feature vectors received for each biological assay and further prediction of class affiliation in relation with the biological assay. The value of mean balanced accuracy obtained during the preliminary testing was about 86%. The results of QSAR models creation based on the homogeneous data sets will be discussed. Classification of HIV-1 RTIs testing methods will be presented.

1. Penazzato M. et al. *Drugs*, 2011, **71(16)**: 2131-49.
2. Kumari G. et al. *Current Pharmaceutical Design*, 2013, **19(10)**: 1767-1783.
3. Ou-Yang S.S. et al. *Acta Pharmacologica Sinica*, 2012, **33(9)**: 1131-40.
4. Liao C. et al. *Future Medicinal Chemistry*, 2010, **2(7)**: 1123-1140.

This work was supported by RFBR grant No. 16-34-60187 and RFBR grant No. 17-54-30015 NIH_a